



THE UNIVERSITY *of* EDINBURGH

Edinburgh Research Explorer

## Bayesian Inference of Atomistic Structure in Functional Materials

**Citation for published version:**

Todorovic, M, Gutmann, M, Corander, J & Rinke, P 2019, 'Bayesian Inference of Atomistic Structure in Functional Materials' npj Computational Materials, vol. 5, 35. DOI: 10.1038/s41524-019-0175-2

**Digital Object Identifier (DOI):**

[10.1038/s41524-019-0175-2](https://doi.org/10.1038/s41524-019-0175-2)

**Link:**

[Link to publication record in Edinburgh Research Explorer](#)

**Document Version:**

Publisher's PDF, also known as Version of record

**Published In:**

npj Computational Materials

**General rights**

Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

**Take down policy**

The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact [openaccess@ed.ac.uk](mailto:openaccess@ed.ac.uk) providing details, and we will remove access to the work immediately and investigate your claim.



## ARTICLE OPEN

## Bayesian inference of atomistic structure in functional materials

Milica Todorović<sup>1</sup>, Michael U. Gutmann<sup>2</sup>, Jukka Corander<sup>3,4</sup> and Patrick Rinke<sup>1</sup>

Tailoring the functional properties of advanced organic/inorganic heterogeneous devices to their intended technological applications requires knowledge and control of the microscopic structure inside the device. Atomistic quantum mechanical simulation methods deliver accurate energies and properties for individual configurations, however, finding the most favourable configurations remains computationally prohibitive. We propose a ‘building block’-based Bayesian Optimisation Structure Search (BOSS) approach for addressing extended organic/inorganic interface problems and demonstrate its feasibility in a molecular surface adsorption study. In BOSS, a Bayesian model identifies material energy landscapes in an accelerated fashion from atomistic configurations sampled during active learning. This allowed us to identify several most favourable molecular adsorption configurations for C<sub>60</sub> on the (101) surface of TiO<sub>2</sub> anatase and clarify the key molecule-surface interactions governing structural assembly. Inferred structures were in good agreement with detailed experimental images of this surface adsorbate, demonstrating good predictive power of BOSS and opening the route towards large-scale surface adsorption studies of molecular aggregates and films.

npj Computational Materials (2019)5:35; <https://doi.org/10.1038/s41524-019-0175-2>

## INTRODUCTION

Frontier technologies are increasingly based on functional hybrid materials—engineered blends of organic molecules and inorganic crystals that harness and enhance the functional properties of both substances to perform specific tasks. Organic/inorganic heterostructures and metal–organic frameworks are key components for smart sensors, membranes and coatings, novel optoelectronic and fuel cell technologies, with further applications in data storage, quantum engineering and nanophotonics on the horizon.<sup>1–8</sup> Despite outstanding component materials, engineering the microscopic structure of complex heterostructures to tailor their properties towards desired functionality remains a fundamental challenge in physics, chemistry and materials science. It means bypassing the pitfalls of interface artifacts, defects and unfavourable self-assembled structures that degrade overall device performance.

Understanding the microscopic structural details of advanced organic/inorganic material blends has emerged as the primary route towards controlling and engineering the functionality of hybrid materials.<sup>2–9</sup> Here, computational studies lead the way,<sup>10,11</sup> since nanoscale experimental measurement techniques frequently lack the necessary atomistic detail, and traditional trial-and-error tests are costly and time-consuming. Ab initio methods like density functional theory (DFT) are especially predictive in simulations of hybrid materials because they accurately describe the delicate interplay of microscopic interactions (e.g., electrostatics, dispersion, bond formation, and charge transfer) that direct structural assembly.<sup>12</sup> DFT maps the atomic structure of a material onto an intrinsic energy, with lower energies indicating more stable material polymorphs. Theoretical structure prediction

methods focus on exploring the resulting configurational phase-space, the potential energy surface (PES).<sup>13,14</sup> Extensive PES sampling by DFT is computationally prohibitive and intractable. In practice it must be reduced to comparing several most-likely structures, which is unreliable in complex materials.

For this reason, hybrid organic/inorganic interfaces present a special challenge for structure search methods. As illustrated in Fig. 1, their PES is complicated by the variety of different morphologies that molecular films can adopt against the solid material. Moreover, the large size of functional molecules means that extensive simulation cells (large length scales) are needed to describe molecular film morphologies, making computations particularly expensive.

To address this structure search problem, we harness the power of AI methods. Recently, AI and machine learning (ML) algorithms were coupled with DFT to approximate the PES<sup>15–17</sup> or improve sampling and accelerate structure prediction in single material clusters and solids.<sup>18–23</sup> Their application to heterostructures is not straightforward, and they may not scale up to required sizes. In some cases, framework setup and the choice of ML parameters was found to affect the results.<sup>15,24</sup> Many schemes rely on large data sets with 1000–10000 sampled points,<sup>25</sup> which are costly to compute. Our ideal method would need to be (i) *efficient* (minimal sampling costs), (ii) *accurate* (both in robust model convergence and DFT chemical accuracy), (iii) *comprehensive* (delivering the entire PES information of global and local minima), (iv) *transferable* (minimal dependence on ML parameters), (v) *versatile* (adaptable to targeting properties, structural prescreening, etc.), (vi) *flexible* (easily combined with other schemes) and (vii) truly *multi-scale* in its scope.

<sup>1</sup>Department of Applied Physics, Aalto University, P.O. Box 11100, Aalto FI-00076, Finland; <sup>2</sup>School of Informatics, University of Edinburgh, 10 Crichton Street, Edinburgh EH8 9AB, UK; <sup>3</sup>Institute of Basic Medical Sciences, University of Oslo, Sognsvannsveien 9, 0372 Oslo, Norway and <sup>4</sup>Department of Mathematics and Statistics, University of Helsinki, P. O. Box 68, Helsinki FI-00014, Finland

Correspondence: Milica Todorović (milica.todorovic@aalto.fi)

Received: 15 June 2018 Accepted: 22 February 2019

Published online: 18 March 2019

Here, we propose an AI-based structure search scheme that is capable of accelerated and unbiased PES computation, and can be extended to large length scales while minimising the amount of configurational sampling. The Bayesian Optimisation Structure Search (BOSS) method, illustrated in Fig. 2a, couples state-of-the-art DFT or quantum chemistry treatment with the BO technique for complex optimisation tasks.

Approximate Bayesian Computation<sup>26</sup> is a class of likelihood-free inference (LFI) methods where data sampling involves complex evaluation. It has recently been combined with BO<sup>27</sup> to accelerate model prediction where data evaluation is also costly. In this work, we adapted the resulting BOLFI scheme<sup>28</sup> to search for minima of the PES in an arbitrary phase space using Gaussian Process (GP) models. BOSS utilises an advanced DFT framework designed for efficient first-principles materials simulations on supercomputer infrastructures.<sup>29</sup> Each data point is a DFT total energy representing an atomistic configuration.

BOSS employs GPs to fit a surrogate PES model to DFT data points, then refines it by acquiring more data points through a smart sampling strategy (see Fig. 2b). The most-likely PES model for given data is the GP posterior mean, which can be traversed by minimisation algorithms to determine all minima and their locations in phase space. The GP posterior variance reflects the lack of confidence in the probabilistic PES model, which vanishes

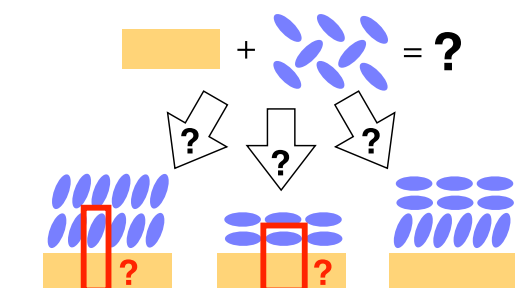
at the data points, and rises in unexplored areas of phase space. In analogy with the 1D example in Fig. 2b, BOSS actively learns every point of the PES in N dimensions and across the defined phase space until convergence is achieved.

Smart sampling of new configurations allows BOSS to make accurate DFT-based predictions despite DFT's computational cost. Our chosen algorithm for sequential acquisition of new energy points combines *exploration* (searching less visited areas) with *exploitation* (searching low-energy areas) to determine the PES global minimum with as few data points as possible. Such a strategy, encapsulated in the exploratory lower confidence bound acquisition function,<sup>27,30</sup> ensures fast determination of the global minimum. We employ an acquisition function that increasingly favours exploration with rising search dimensionality and iteration step.<sup>28</sup>

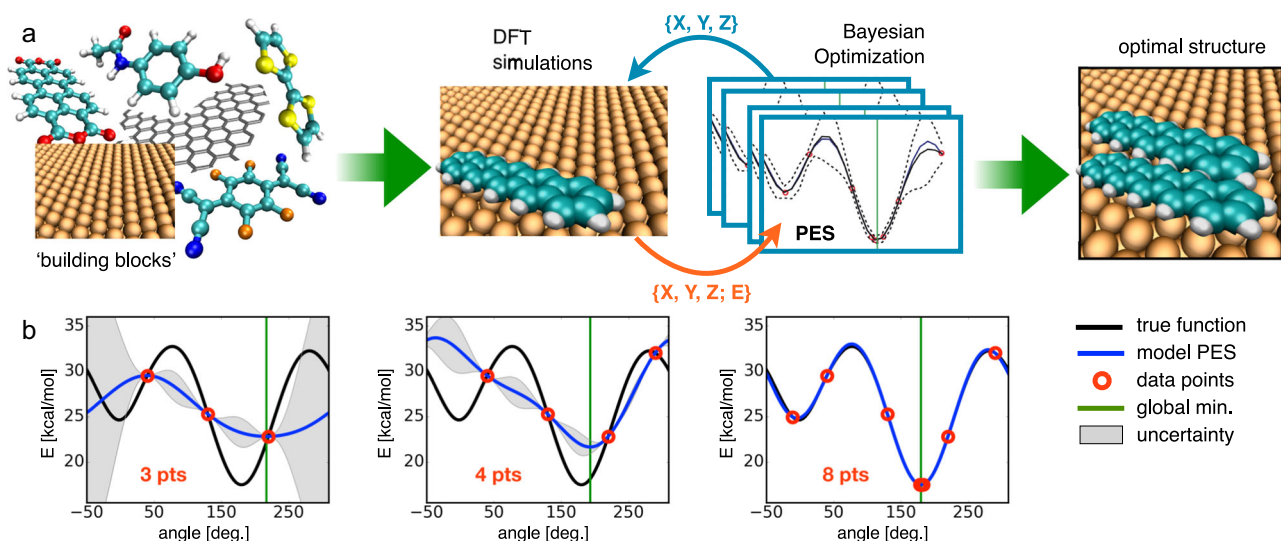
A common feature of structure search in complex heterogeneous materials is the presence of rigid organic and inorganic structures, (aromatic rings and functional groups), where structure change is confined to small bond adjustments, without bond rearrangements. To expedite structure search over large numbers of atoms, we follow other schemes<sup>31,32</sup> and fix these internal components of the material to rigid 'building block' components. Such an approach is suitable to describe molecular physisorption and some chemisorption via anchoring groups, both common at hybrid interfaces. The choice of building blocks is motivated by chemical rules, and expedites the search by confining it to configurational phase space, instead of full chemical phase space.

In the long-term, BOSS can be used to predict the structure of organic/inorganic interfaces by identifying the most stable organic thin-film morphologies on inorganic substrates. The procedure is illustrated in Fig. 2a: once the simulation 'building blocks' are identified, the learning would progress from single adsorbates to molecular aggregates and monolayers. While some methods acquire single-adsorption configurations by intuition and focus on complex lattice-based film morphology search,<sup>33,34</sup> we aim to treat both the molecular adsorbates and aggregates within the BOSS framework by increasing search degrees of freedom.

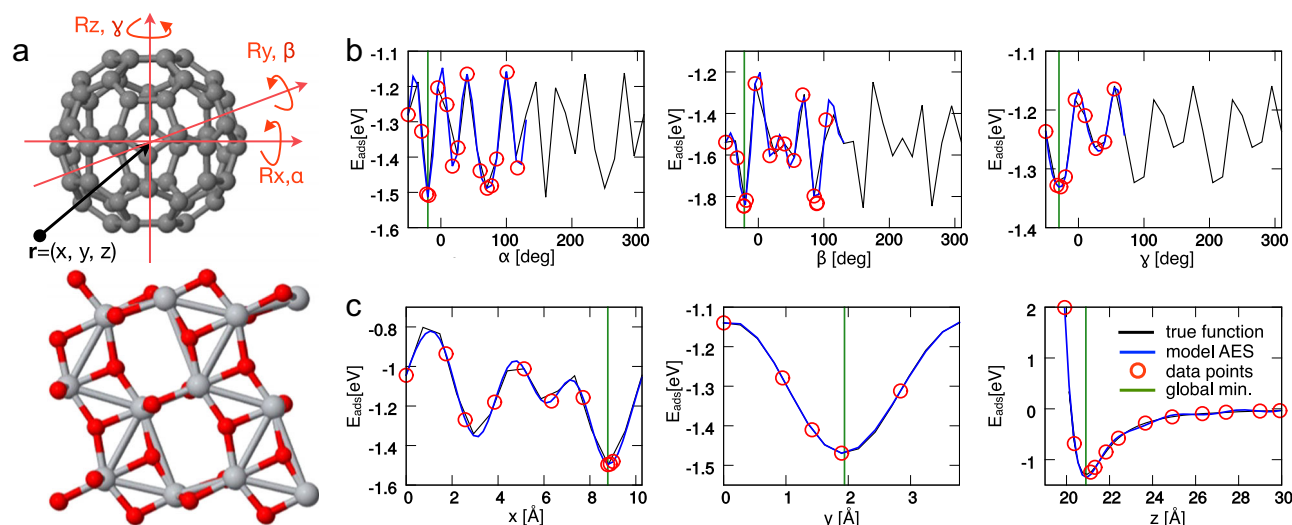
Learning the individual molecule-surface interactions and structure is a *key step*, which is demonstrated here by applying BOSS to infer the structure of a single-molecular surface adsorbate. In this manuscript, we conducted a structural study



**Fig. 1** Inside devices, various thin-film morphologies composed from organic molecules (blue oval) may be formed at the interface with a crystalline solid (yellow). In the first step towards a large-scale Bayesian structure search of monolayer morphologies, we focus on inferring the configuration of a single-molecule adsorbate (shown in red)



**Fig. 2** Illustration of a typical BOSS application. **a** Schematic of key steps in BOSS structure search at the inorganic surface: from the choice of materials and building blocks, through selection of the BO degrees of freedom and the iterative optimisation, towards the inferred individual adsorbate and thin-film structures. **b** Example of BOSS iterative inference of a simple one-dimensional (1D) PES featuring a global and local minimum. The GP native uncertainty (grey areas) facilitates exploratory data sampling. The global minimum location and the entire PES are learned in eight data acquisitions



**Fig. 3** BOSS applied to the  $C_{60}/TiO_2$  adsorption problem. **a** Atomistic model of  $C_{60}$  on the (101) surface of  $TiO_2$  anatase in the reference configuration, with the energetically dominant degrees of freedom for the molecule indicated in black (translational motion) and red (rotational motion). **b** Comparison of the converged 1D AES with the true function for all rotational variables. **c** Comparison of the converged 1D AES with the true function for all translational degrees of freedom. Learning in **b** and **c** was initialised with five quasi-random points and the models converged in up to seven BO acquisitions. 1D searches were carried out with the other variables fixed to reference values, as illustrated in **a** and described in the Methods section

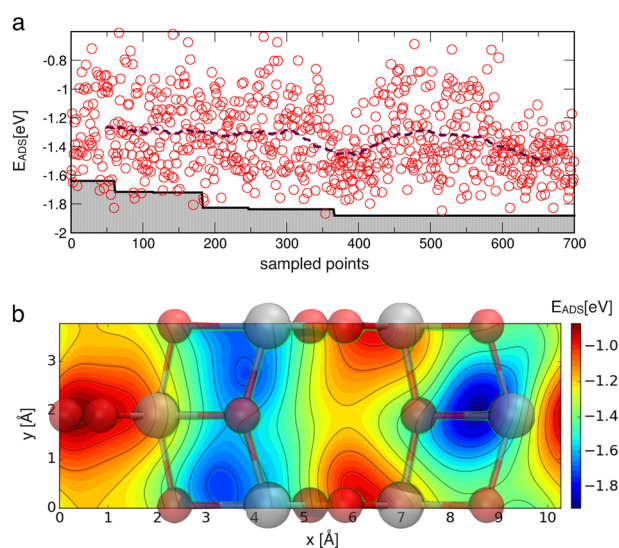
of a fullerene molecule on the (101) surface of  $TiO_2$  anatase. Both are functional materials frequently employed in organic optoelectronics.<sup>35–37</sup> To verify AI predictions, inferred structures were compared to the high-resolution atomic force microscopy (AFM) images<sup>38</sup> of this surface system.

## RESULTS

The atomistic simulation model of  $C_{60}$  on the  $TiO_2$  anatase surface is presented in Fig. 3a. The surface slab and fullerene cage were defined as building blocks. Stable molecular adsorbate structures are the atomistic configurations that minimise the adsorption energy, so BOSS was set to learn the adsorption energy surface (AES). The adsorption energy depends on the molecular position above the surface, represented by the molecular centre of mass  $r = [x, y, z]$ , and its orientation towards the surface. The latter was described by angles of rotation  $\alpha$ ,  $\beta$  and  $\gamma$  with respect to Cartesian axes of rotation  $R_x$ ,  $R_y$  and  $R_z$ , respectively.

The full AES is a 6-dimensional (6D) function of rotational and translational degrees of freedom  $E_{AES} = E(\alpha, \beta, \gamma, x, y, z)$ . In Fig. 3b, c, we present a BOSS investigation into each of these variables separately, which revealed the approximate AES variation from  $-1$  eV to  $-2$  eV. The  $z$  variable was found to produce only a vertical shift in the adsorption energy. The location of the minima in other dimensions did not change with  $z$ , so we fixed it and carried out the full adsorption site BOSS search in 5D.

Figure 4a illustrates the refinement of the predicted 5D global minimum with iterative configurational sampling. The lowest observed adsorption energy  $E_{ADS}$  (computed from BOSS-predicted global minimum locations) converged after 370 sampled configurations to a value of  $-1.88$  eV. Improvement of the global minimum prediction could be correlated to instances of visiting low-energy configurations, chosen strategically from a vast 5D phase space. However, most model refinement proceeded with input from less relevant configurations, on average in the region  $0.5$  eV above the predicted global minimum (after 400 iterations, the average acquisitions shifted to lower values, suggesting that the model is exploring near local minima). A physically meaningful 5D model of the  $E_{ADS}$  landscape (consistent with the symmetries of the DFT simulations) converged after 670 data acquisitions.



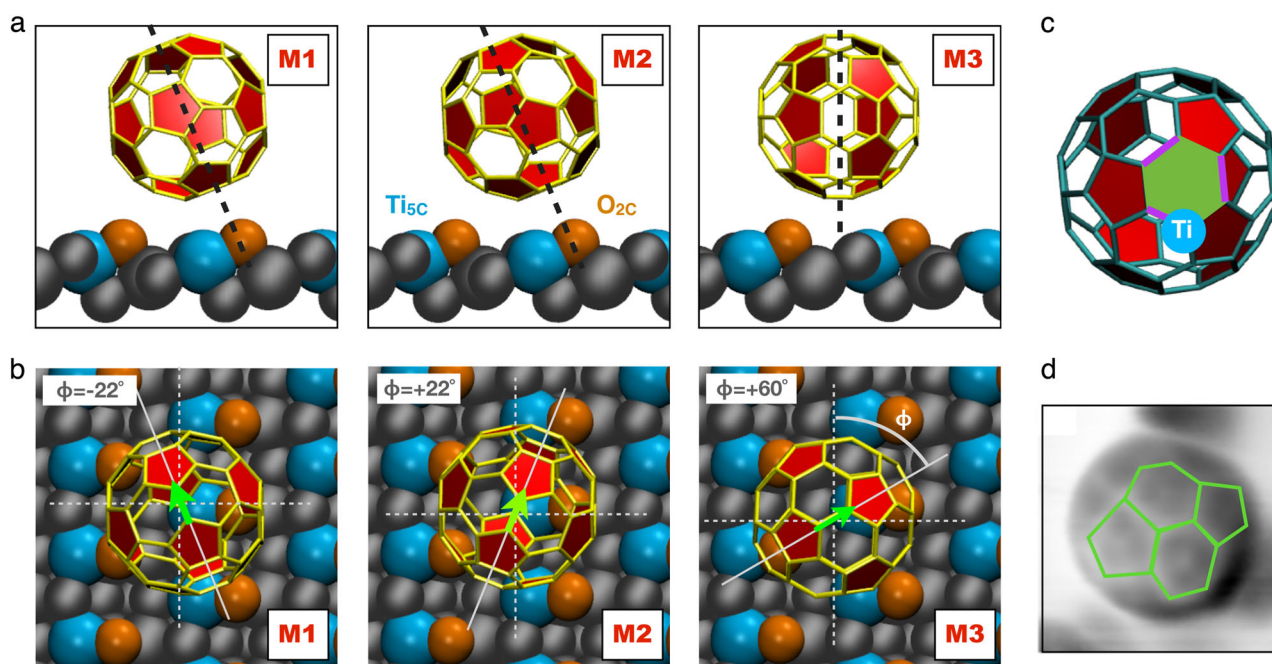
**Fig. 4** 5D BOSS search results. **a** Convergence of the  $E_{ADS}$  computed from the BOSS global minimum prediction during active learning (black line). The accuracy of the inferred result improved with strategic 5D configurational sampling (red data points, with running average shown in dashed line). **b** 2D cross-section of the 5D BOSS search illustrating  $x$ - $y$  molecular translation, extracted at the 5D global minimum after 700 data acquisitions. The first  $TiO_2$  surface layer shown in overlay reveals the correlation between the global minimum (deep blue) and the  $Ti_{5c}$  surface site (grey atom)

Figure 4b shows the  $x$ - $y$  cross-section extracted from the 5D model at iteration 700. The AES landscape correlates well with the two sloping terraces of the  $TiO_2$  surface. BOSS typically finds the global minimum quickly, while more data is needed to refine the entire PES model.

## Chemical insight from AI models

The chemically natural assignment of ‘building blocks’ means that resulting energy curves lend themselves readily to human interpretation. Already the preliminary 1D BOSS simulations





**Fig. 5** Verifying BOSS-predicted structures. **a** Side view of the three lowest energy adsorption configurations M1, M2 and M3 obtained by full structural relaxation from BOSS-predicted minima. Pentagonal facets of C<sub>60</sub> are coloured in red for visual distinction and the symmetry axis for molecular rotation is indicated by the black dashed line. Reactive under-coordinated atoms on the surface are shown in blue (Ti<sub>5c</sub>) and orange (O<sub>2c</sub>) to highlight molecular registry on the surface. **b** Top view of the three lowest energy adsorption configurations M1, M2 and M3. The green arrow illustrates the direction of the typical oval feature observed in all three structures, along the bond between two hexagons. The angle  $\phi$  denotes the orientation of the bond with respect to the [010] crystallographic direction. **c** Underside of the C<sub>60</sub> cage directly above the Ti<sub>5c</sub> surface binding site. Molecular binding is facilitated by the hexagonal facet (green) and the nearby C<sub>h</sub>-C<sub>h</sub> bond (purple). **d** Frequency shift response submolecular AFM image of C<sub>60</sub> on the (101) surface of TiO<sub>2</sub> anatase, with green lines indicating the top facets. (Adapted with permission from Moreno et al.<sup>38</sup>. Copyright (2015) American Chemical Society)

revealed a wealth of information about the binding and structure at the C<sub>60</sub>/TiO<sub>2</sub> interface. This helped us to determine the key binding sites on both the molecule and the surface.

Translations of the molecule across the surface produced slowly-varying energies with few minima (Fig. 3c), closely reflecting anatase corrugation. The surface adsorption site was the Ti<sub>5c</sub> or the O<sub>3c</sub> one, depending on molecular orientation. Molecular rotation gave rise to complex fast-varying AES curves with multiple deep minima (Fig. 3b), as expected from the high symmetry of the C<sub>60</sub> cage. By analysing 1D global minima in  $\beta$  (−1.85 eV) and  $\alpha$  (−1.50 eV), we determined the active sites on the molecular cage to be the hexagonal facet and the C<sub>h</sub>-C<sub>h</sub> bond between them (respectively).

These findings are consistent with the global minimum structure inferred in the 5D AI search. Molecular rotation was the energetically dominant factor for surface adsorption. The global minimum orientation of the physisorbed C<sub>60</sub> cage featured the hexagonal facet roughly parallel to the anatase terrace. The optimal surface adsorption site was located above the under-coordinated Ti<sub>5c</sub> surface atom, the site identified as most reactive on this surface by earlier studies of small adsorbates.<sup>39,40</sup>

#### Verifying BOSS-predicted structures

The BOSS AES search converged with a global energy minimum of  $E_{\text{BOSS}} = -1.9$  eV within the constraint of the structural ‘building blocks’. To verify the quality of the prediction, we removed this approximation and allowed all degrees of freedom to relax in DFT. The structure remained the same, with the overall shift in all atomic positions described by a nominal root mean squared distance of 0.19 Å. The resulting global minimum  $E_{\text{GL}} = -2.0$  eV (0.1 eV below the AI value) and the minimal change in bond

lengths (below 0.01 Å) indicated that the ‘building block’ approximation was appropriate in this case study.

Next, we compared predicted structures with experimental observations. In addition to the global minimum, we considered the nearest six unique local minima located by BOSS within a 0.1 eV energy window from the 5D global minimum. This allowed us to compare a range of low-energy adsorption configurations with experimental structures, where molecules evaporated onto a hot surface may have acquired similar thermal energy. After seven full structural optimisations, all structures were reduced to one of three different configurations in Fig. 5a.

The M1 adsorption geometry was qualitatively identified as the BOSS-predicted 5D global minimum, with M2 as its degenerate mirror image (by 180° rotation about the axis perpendicular to the anatase terrace). Both the hexagonal C<sub>60</sub> facet and a nearby C<sub>h</sub>-C<sub>h</sub> bond approached the surface (see Fig. 5c). The more symmetric M3 configuration in Fig. 5a was the only local energy minimum found, with an energy of  $E_{\text{loc}} = -1.93$  eV. The 5D BOSS search thus led us to non-symmetric low-energy configurations stabilised by competing interactions. Any symmetric initial guess structure would likely have failed to reach the deeper energy minimum during structure optimisation.

An AFM experimental image with sub-molecular resolution of C<sub>60</sub> on the surface of TiO<sub>2</sub> anatase is presented in Fig. 5d. For comparison, we considered the top-down view of the three absorption configurations in Fig. 5b. An elliptical feature with two hexagonal and two pentagonal facets is visible at the top of the molecules. We defined the direction of the feature along the bond separating the two hexagons (the long axis of the ellipse) and computed its orientation with respect to the [010] crystallographic direction to serve as an identification fingerprint. A similar elliptical feature in the AFM image points to good qualitative agreement between experiment and theory. The M1 and M2

molecular structures are topped by a central C atom at the edge of the C–C bond, just like in the experimental image (other BOSS local minima structures were topped by a C–C bond, and we found none topped by a planar facet as in Fig. 2a). The lack of substrate information made it difficult to conclusively identify the experimental structural fingerprint.

### Sampling efficiency

To evaluate the efficiency of BO in structure search, we consider the number of sampled configurations required to converge the global minimum prediction, and later, the AES landscape model. We are not aware of other structure search methods that could provide a comparison. Instead, we compare our method against conventional techniques for determining molecular adsorbate structures: grid-based sampling and human intuition paired with geometry optimisation.

BOSS was quick to locate the global minimum in all test cases. 1D the 2D global minima were identified after 10 and 30 visited configurations, respectively. Predictions converged with 150–300 data points in various 3D–4D cases, and 370 in the 5D case. This is a remarkably low computational effort given the vast search space.

In computing the energy landscapes, the number of required data points rose with search dimensionality as well as the complexity of the search (number of minima). All the preliminary 1D models in Fig. 2 required less than 12 data points to converge, at least twice as fast as the grid-based computation of the true energy function with the same resolution. In 2D BOSS tests, the  $x$ – $y$  landscape was obtained after 45 data points (one minimum), but the more complex  $\alpha$ – $\beta$  one required 90 acquisitions (16 minima). The same resolution in the  $\alpha$ – $\beta$  AES would require some 500 acquisitions with grid-based methods.

Grid searches become impractical beyond 3D, whereas BOSS produced good quality AES models also in 3D and 4D simulations (not shown here). These could be sliced in 2D to facilitate the interpretation of the molecule-surface interactions. The many reactive sites of the symmetric  $C_{60}$  cage presented a major challenge for learning the entire AES in 5D, yet BOSS resolved it with only 700 data points. In an intuition-led force minimisation adsorption study, such a computational effort would yield optimised structures from 20 to 30 different initial guess configurations (assuming that every structure relaxation converges in 20 to 30 single-point DFT calculations). We might choose the best candidate between them, with no possibility of checking if any unknown lower energy structures exist. With AI, 700 data points deliver the optimal configuration across the entire phase space, and additionally, all the local minima and the barriers between them.

## DISCUSSION

We developed an AI-based structure search technique for complex materials that is in line with our ideal methodology described in the Introduction. The BOSS scheme is certainly (i) *efficient* and (ii) *accurate* in finding the global minimum in 6D (350 DFT evaluations) compared to the traditional structure search strategy. Ultimately, fewer than 100 evaluations would be desirable and further method development (accounting for energy gradients and material symmetries) should considerably speed up the inference. The (iii) *comprehensive* nature of the scheme (global and local minima available) comes at the cost of further computational effort, but the type and the amount of information obtained by inferring the entire energy landscape is not available from other structure search methods. Designing methodology to extract minimum energy paths from N-dimensional energy landscapes would make our scheme even more comprehensive.

Our case study indicates that BOSS is a (iv) *transferable* technique since it inferred both fast and slow varying energy functions by successfully converging parameters on the fly (Fig. 3). Nevertheless, further work on diverse test cases is needed to better characterise method transferability. BOSS is designed for general degrees of freedom, which facilitates (vi) *flexibility* in workflows with other ML-based structure search techniques. It could be employed for global conformer search of small molecules before these are inserted into the Gator genetic algorithm scheme for organic crystal structure search,<sup>41</sup> or for determining adsorption structures of individual molecules to be employed in registry-based film morphology studies.<sup>33</sup>

BOSS is certainly (v) *versatile*, since multiple energetic and electronic structure properties are available from each DFT acquisition. Consequently, the inference could be targeted to optimal properties or multi-target objectives instead. It appears straightforward to extend BOSS to (vii) *multi-scale* molecular film simulations, but method performance with increasing dimensionality requires thorough characterisation. Bayesian optimisation scaled better than expected up to 6D (not exponential) on account of periodic kernels employed, and in future work we plan to carry out a quantitative analysis of dimensionality scaling for different tests cases. In our ultimate goal of predicting film formation and morphology we have achieved the first step of having an efficient method for individual molecules on surfaces. We can now build on this to extend BOSS to higher dimensionality (i.e., more than one molecule) or couple it to multi-scale schemes tailored for molecular ensembles.

In summary, we proposed a novel structure search scheme that combines a smart AI sampling strategy and a natural ‘building blocks’ representation with accurate quantum mechanical calculations. As the first step in targeting the structure of large-scale molecular films and organic/inorganic interfaces, we employed it to learn the adsorption structure of a single molecule:  $C_{60}$  on the (101) surface of  $TiO_2$  anatase.

The BOSS approach facilitated a computationally tractable study of molecular adsorption as a function of key degrees of freedom, molecular registry and orientation. The correct global minimum, verified against fully optimised structures, was located in multi-dimensional phase space with considerable efficiency. Structures based on BOSS-inferred models were in good agreement with high-resolution experimental images of this material. Additional sampling allowed us to compute multi-dimensional AES energy landscapes, with meaningful local minima and energy barriers between them. The resulting chemical insight into the molecule-surface interactions helped us interpret the predicted adsorption structures. Future model refinement could be made more robust by using GP prior belief functions, different GP kernels and by explicitly accounting for material symmetry.

The ‘building block’ approach served very well for  $C_{60}$  adsorbed on  $TiO_2$  anatase, and will allow us to readily extend our approach to multi-scale simulations. In short, our BOSS scheme delivers on many fronts in a successful study of molecular surface adsorption and further work will see it applied to more complex configurational studies of surface-supported molecular aggregates and films.

## METHODS

### AI software

BOLFI based on the *gpml* package<sup>42</sup> was implemented in a serial MATLAB code, which was interfaced with the total energy simulation method. The knowledge about the PES was encoded in the GP, characterised by the GP posterior mean (PES model) and variance functions. The posterior variance supplied a measure of uncertainty on the probabilistic model. We employed a non-isotropic standard periodic GP kernel to account for periodic boundary conditions. Initial sampled data points were selected by a Sobol quasi-random sequence generator, upon which the BO process

was initialised. The scheme features only two hyperparameters, which are also learned on-the-fly. The GP model and its hyperparameters were updated every ten acquisitions until convergence. We analysed the standard deviation on the GP posterior mean: this error remained 0.1 eV on average, or 6% of the energy minimum. We also monitored model quality by noting the convergence of local and global minima, as well as qualitatively checking model cross-sections for the expected symmetries of the atomic model.

### First-principles calculations

We performed all configurational sampling with the all-electron DFT code FHI-aims.<sup>29</sup> Simulations were carried out with converged Tier 2 basis sets free of *g* and *h* functions, and the PBE exchange-correlation functional<sup>43</sup> augmented with van der Waals correction terms.<sup>44</sup> Relativistic corrections accounted for heavy elements. *Light* grids with  $\Gamma$ -point reciprocal space sampling was employed to build the PES model. Global minima structures were verified with *tight* grids and a  $2 \times 2 \times 1$  k-point mesh, which lead to the same geometries, but reduced the adsorption energy by 0.3 eV. With the efficient code parallelisation,<sup>45</sup> a single-acquisition calculation on 168 atoms required 10 min on 120 central processing units. The (101) TiO<sub>2</sub> anatase surface slab featured three typical trilayers in a  $10.27 \text{ \AA} \times 11.36 \text{ \AA} \times 52.77 \text{ \AA}$  periodic unit cell, exposing a  $1 \times 3$  unit cell surface area.<sup>46</sup> Molecular adsorption energies converged with three trilayers; the lowest two trilayers were kept fixed during structural optimisations.

To define the boundaries of BOSS search phase space, we relied on the surface and molecule symmetry and periodicity. Molecular registry search space was limited to the smallest periodically repeating surface unit  $10.27 \text{ \AA} \times 3.78 \text{ \AA}$  and informed by this periodicity. The non-periodic *z* variable search was conducted  $10 \text{ \AA}$  in height from the  $1.5 \text{ \AA}$  closest surface approach. The high symmetry of the C<sub>60</sub> cage was broken by the asymmetric surface features, allowing us to take limited advantage of molecular symmetry. Molecular orientation search was conducted in minimal unique periods of 180 for  $\alpha$  and  $\beta$  angles, and 120 for the  $\gamma$  angle, exploiting the symmetry of the C<sub>60</sub> cage. The local minimum reference configuration in Fig. 2a, was employed to initialise the BOSS search and set the values for fixed variables when required: (*x*, *y*) to (0,0) coordinates in Fig. 4b. (approximately the mid-point between two O<sub>2c</sub> sites on the surface), *z* =  $2.2 \text{ \AA}$  above the surface, and the angles to (0, 0, 0) as indicated by Fig. 2a.

### DATA AVAILABILITY

The dataset generated during the study is available at an online repository.<sup>47</sup>

### ACKNOWLEDGEMENTS

This work was supported by the Academy of Finland through Project Nos. 251748, 284621 and 316601, and also through the European Union's Horizon 2020 research and innovation programme under Grant agreement No. 676580 with The Novel Materials Discovery (NOMAD) Laboratory, a European Center of Excellence. J.C. was funded by the ERC grant no. 742158. Computer time was provided by the Centre for Scientific Computing (CSC, Finland) at the Taito supercomputer.

### AUTHOR CONTRIBUTIONS

P.R., J.C., M.U.G. and M.T. designed the research. M.U.G. contributed BO routines, M.U.G. and M.T. produced the BOSS code. M.T. performed the research, analysed the data and wrote the paper. All the authors read the manuscript, commented the contents and agreed with the publication of the results. J.C. and P.R. contributed equally to this work.

### ADDITIONAL INFORMATION

**Competing interests:** The authors declare no competing interests.

**Publisher's note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

### REFERENCES

- Theobald, J. A., Oxtoby, N. S., Phillips, M. A., Champness, N. R. & Beton, P. H. Controlling molecular deposition and layer structure with supramolecular surface assemblies. *Nature* **424**, 1029–1031 (2003).
- Barth, J. V., Costantini, G. & Kern, K. Engineering atomic and molecular nanostructures at surfaces. *Nature* **437**, 671–679 (2005).
- Grill, L. et al. Nano-architectures by covalent assembly of molecular building blocks. *Nat. Nano* **2**, 687–691 (2007).
- Schlesinger, R. et al. Efficient light emission from inorganic and organic semiconductor hybrid structures by energy-level tuning. *Nat. Commun.* **6**, 6754 (2015).
- Denny, M. S. Jr., Moreton, J. C., Benz, L. & Cohen, S. M. Metal–organic frameworks for membrane-based separations. *Nat. Rev. Mater.* **1**, 16078 (2016).
- Huang, N., Wang, P. & Jiang, D. Covalent organic frameworks: a materials platform for structural and functional designs. *Nat. Rev. Mater.* **1**, 16068 (2016).
- Yan-fang, G. et al. STM probing the supramolecular coordination chemistry on solid surface: Structure, dynamic, and reactivity. *Coord. Chem. Rev.* **337**, 145–177 (2017).
- Song, Y. et al. Self-assembly and local manipulation of Au-pyridyl coordination networks on metal surfaces. *Chem. Phys. Chem.* **18**, 2088–2093 (2017).
- Howarth, A. J. et al. Chemical, thermal and mechanical stabilities of metal–organic frameworks. *Nat. Rev. Mater.* **1**, 15018 (2016).
- von Lilienfeld, O. A. *Towards the Computational Design of Compounds from First Principles* (Springer International Publishing, Switzerland, 2014).
- Curtarolo, S. et al. The high-throughput highway to computational materials design. *Nat. Mater.* **12**, 191–201 (2013).
- Jones, R. O. Density functional theory: its origins, rise to prominence, and future. *Rev. Mod. Phys.* **87**, 897 (2015).
- Kirkpatrick, S., Gelatt, C. D. & Vecchi, M. P. Optimization by simulated annealing. *Science* **220**, 671–680 (1983).
- Goedecker, S., Hellmann, W. & Lenosky, T. Global minimum determination of the Born–Oppenheimer surface within density functional theory. *Phys. Rev. Lett.* **95**, 055501 (2005).
- Behler, J. Representing potential energy surfaces by high-dimensional neural network potentials. *J. Phys.* **26**, 183001 (2014).
- Bartok, A. P., Payne, M. C., Kondor, R. & Csanyi, G. Gaussian approximation potentials: the accuracy of quantum mechanics, without the electrons. *Phys. Rev. Lett.* **104**, 136403 (2010).
- Li, Z., Kermode, J. R. & De Vita, A. Molecular dynamics with on-the-fly machine learning of quantum-mechanical forces. *Phys. Rev. Lett.* **114**, 096405 (2015).
- D'Avezac, M. & Zunger, A. Identifying the minimum-energy atomic configuration on a lattice: Lamarckian twist on Darwinian evolution. *Phys. Rev. B* **78**, 064102 (2008).
- Wang, Y., Lv, J., Zhu, L. & Ma, Y. CALYPSO: a method for crystal structure prediction. *Comput. Phys. Commun.* **183**, 2063–2070 (2012).
- Nelson, L. J., Hart, G. L. W., Zhou, F. & Ozolins, V. Compressive sensing as a paradigm for building physics models. *Phys. Rev. B* **87**, 035125 (2013).
- Bhattacharya, S., Levchenko, S. V., Ghiringhelli, L. M. & Scheffer, M. Stability and metastability of clusters in a reactive atmosphere: theoretical evidence for unexpected stoichiometries of MgMOx. *Phys. Rev. Lett.* **111**, 135501 (2013).
- Kiyohara, S., Oda, H., Tsuda, K. & Mizoguchi, T. Acceleration of stable interface structure searching using a kriging approach. *Jpn. J. Appl. Phys.* **55**, 045502 (2016).
- Xue, D. et al. Accelerated search for materials with targeted properties by adaptive design. *Nat. Commun.* **7**, 11241 (2016).
- Bhattacharya, S., Levchenko, S. V., Ghiringhelli, L. M. & Scheffer, M. Efficient ab initio schemes for finding thermodynamically stable and metastable atomic structures: benchmark of cascade genetic algorithms. *New J. Phys.* **16**, 123016 (2014).
- Ueno, T., Rhone, T. D., Hou, Z., Mizoguchi, T. & Tsuda, K. COMBO: an efficient Bayesian optimization library for materials science. *Mater. Discov.* **4**, 18–21 (2016).
- Lintusaari, J., Gutmann, M. U., Dutta, R., Kaski, S. & Corander, J. Fundamentals and recent developments in approximate Bayesian computation. *Syst. Biol.* **66**, e66–e82 (2017).
- Brochu, E., Cora, V. M., & de Freitas, N. A Tutorial on Bayesian Optimization of Expensive Cost Functions, with Application to Active User Modeling and Hierarchical Reinforcement Learning Preprint at <https://arxiv.org/abs/1012.2599> (2010).
- Gutmann, M. U. & Corander, J. Bayesian optimization for likelihood-free inference of simulator-based statistical models. *J. Mach. Learn. Res.* **17**, 1–47 (2016).
- Blum, V. et al. Ab initio molecular simulations with numeric atom-centered orbitals. *Comput. Phys. Commun.* **180**, 2175–2196 (2009).
- Cox, D. D. & John, S. A statistical method for global optimization. In: *Proceedings of the IEEE Conference on Systems, Man and Cybernetics*, pp. 1241–1246, Institute of Electrical and Electronics Engineers (IEEE), Chicago IL, (1992).
- Behler, J., Lorenz, S. & Reuter, K. Representing molecule-surface interactions with symmetry-adapted neural networks. *J. Chem. Phys.* **127**, 014705–014705 (2007).
- Oganov, A. R. & Glass, C. W. Crystal structure prediction using ab initio evolutionary techniques: principles and applications. *J. Chem. Phys.* **124**, 244704 (2006).
- Obersteiner, V., Scherbela, M., Hörmann, L., Wegner, D. & Hofmann, O. T. Structure prediction for surface-induced phases of organic monolayers: overcoming the combinatorial bottleneck. *Nano Lett.* **17**, 4453–4460 (2017).
- Packwood, D. M., Han, P. & Hitosugi, T. Chemical and entropic control on the molecular self-assembly process, *Nature. Nat. Commun.* **8**, 14463 (2017).



35. Grätzel, M. et al. Solid-state dye-sensitized mesoporous TiO<sub>2</sub> solar cells with high photonto-electron conversion efficiencies. *Nature* **395**, 583–585 (1998).
36. Yoo, S. et al. Analysis of improved photovoltaic properties of pentacene/C60 organic solar cells: effects of exciton blocking layer thickness and thermal annealing. *Solid State Electron.* **51**, 1367–1375 (2007).
37. Cheyns, D., Gommans, H., Odijk, M., Poortmans, J. & Heremans, P. Stacked organic solar cells based on pentacene and C60. *Sol. Energy Mater. Sol. Cells* **91**, 399–404 (2007).
38. Moreno, C., Stetsovych, O., Shimizu, T. K. & Custance, Ó. Imaging three-dimensional surface objects with submolecular resolution by atomic force microscopy. *Nano Lett.* **15**, 2257–2262 (2015).
39. Tilocca, A. & Selloni, A. Methanol adsorption and reactivity on clean and hydroxylated anatase (101) surfaces. *J. Phys. Chem. B* **108**, 19314–19319 (2004).
40. He, Y., Tilocca, A., Dulub, O., Selloni, A. & Diebold, U. Local ordering and electronic signatures of submonolayer water on anatase TiO<sub>2</sub>(101). *Nat. Mater.* **8**, 585–589 (2009).
41. Curtis, F. et al. GAtor: a first-principles genetic algorithm for molecular crystal structure prediction. *J. Chem. Theory Comput.* **14**, 2246–2264 (2018).
42. Rasmussen, C. E. & Williams, C. K. I. *Gaussian Processes for Machine Learning*, 2nd edn. (MIT Press, Cambridge, MA, 2006).
43. Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.* **77**, 3865–3868 (1996).
44. Tkatchenko, A. & Scheffler, M. Accurate molecular Van der Waals interactions from ground-state electron density and free-atom reference data. *Phys. Rev. Lett.* **102**, 73005 (2009).
45. Marek, A., Blum, V., Johanni, R. & Havu, V. The ELPA library: scalable parallel eigenvalue solutions for electronic structure theory and computational science. *J. Phys.* **26**, 213201 (2014).
46. Stetsovych, O. et al. Atomic species identification at the (101) anatase surface by simultaneous scanning tunnelling and atomic force microscopy. *Nat. Commun.* **6**, 7265 (2015).
47. Todorovic, M., Gutmann, M. U., Corander, J. & Rinke, P. Bayesian Inference of Atomistic Structure in Functional Materials. Zenodo, <https://doi.org/10.5281/zenodo.2565933> (2019).



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2019